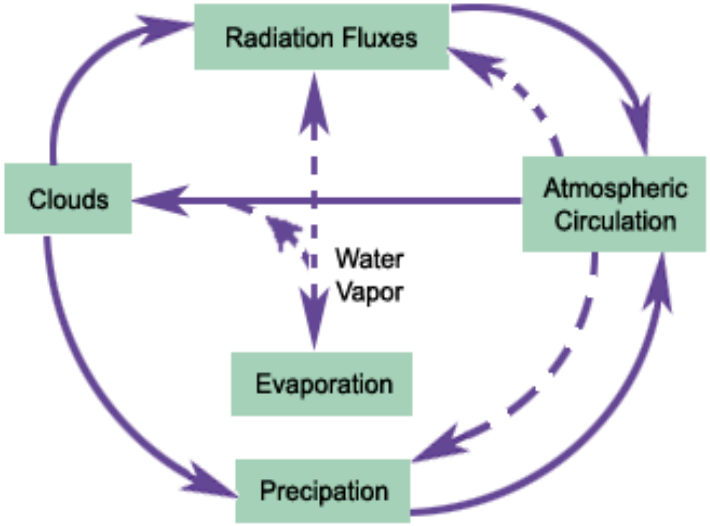
Revealing Relationships among Relevant Climate Variables with Information Theory

Kevin H. Knuth[1], Anthony Gotera[1], Charles T. Curry[1],
Karen A. Huyser[1], Kevin R. Wheeler[1], William B. Rossow[2]

1. Intelligent Systems Division, NASA Ames Res. Center, Moffett Field CA 94035
2. NASA Goddard Institute for Space Studies, New York NY 10025

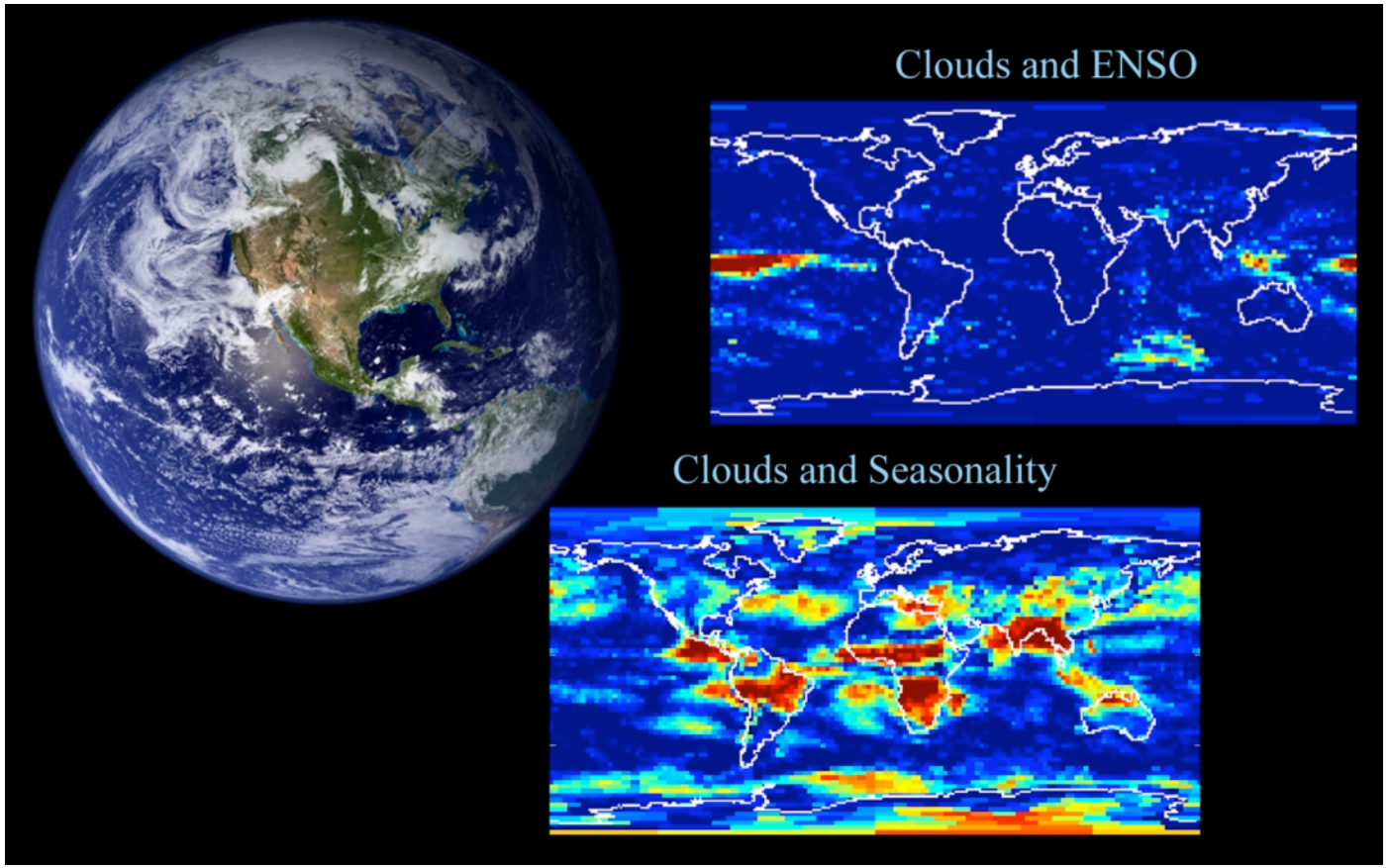
Forcings and Feedback



NASA's goal is to understand the observed Earth climate variability, and determine and predict the climate's response to both natural and human-induced forcing.

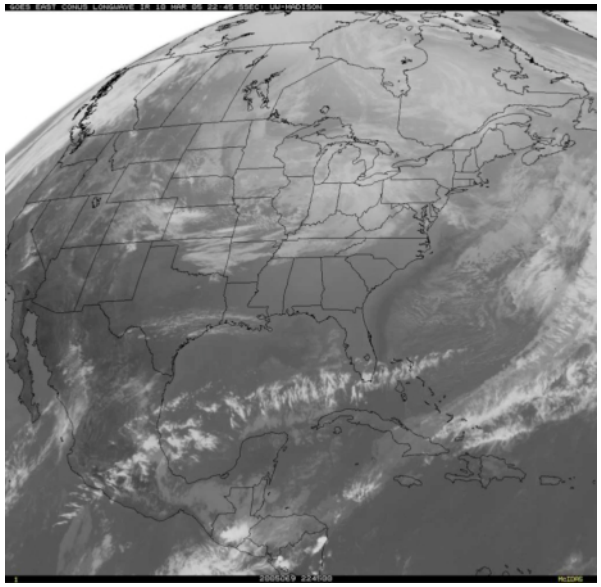
The basic idea is that changes in one climate subsystem will cause or force responses in other subsystems. These responses in turn feed back to force other subsystems.

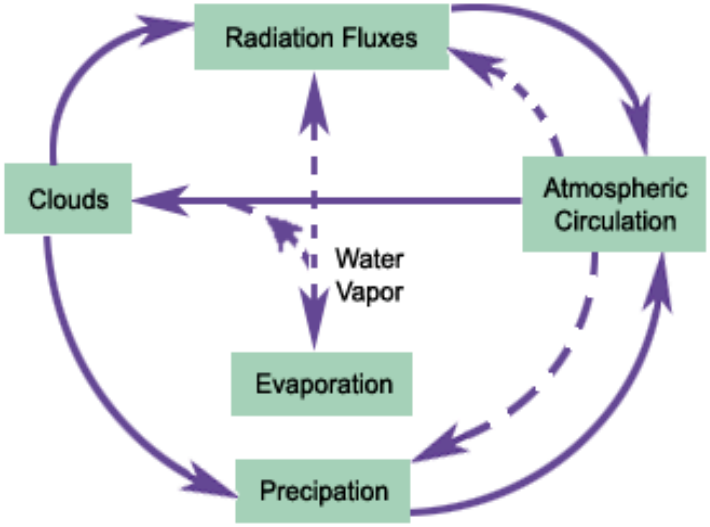
Relevant Variables



One of the greatest difficulties in this field is the identification of the **RELEVANT VARIABLES**.

Causal Interactions





Once relevant variables are identified, one can begin examining their **CAUSAL INTERACTIONS**, which implement forcings and feedbacks.

However, these calculations are useless without **error bars** that indicate our degree of **uncertainty**. Much of the work we propose to do is aimed toward quantifying our uncertainties.

The Underlying Issues

How does one identify relevant variables?

How does one identify, characterize and quantify causal interactions?

How do we quantify the uncertainty in our results?

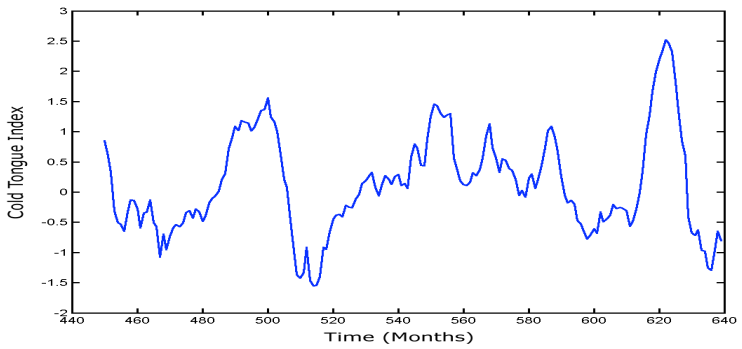
The Underlying Issues

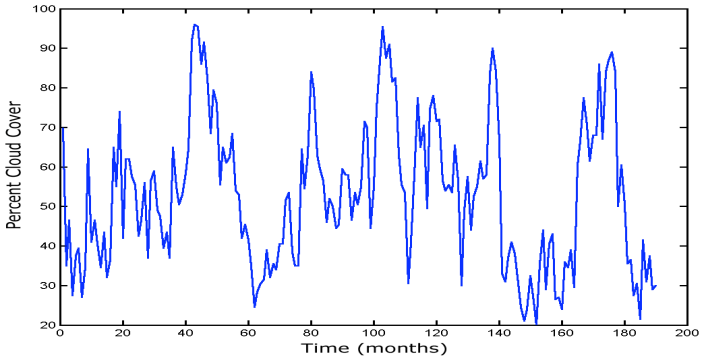
How does one identify relevant variables?

How does one identify, characterize and quantify causal interactions?

How do we quantify the uncertainty in our results?

When the data look like this…





Our Efforts

Research Team

Kevin H. Knuth, PI
NASA Ames

William B. Rossow, co-I
NASA GISS

Anthony Gotera
Cal State East Bay
Mathematics

Karen A. Huyser
Stanford University
Electrical Engineering

Charles R. Curry
UC Santa Cruz
Applied Mathematics

Outline

Information Theory

Modeling Probability Densities

Entropy Estimation

Results

Next Steps

Entropy

We can characterize the behavior of a system X by looking at the set of states the system visits as it evolves in time. If a state is visited rarely, we would be surprised to find the system there. We can express the expectation (or lack of expectation) to find the system in state x in terms of the probability that it can be found in that state, $p(x)$, by

$$h(x) = \log \frac{1}{p(x)}$$

This quantity is often called the **surprise**, since it is large for improbable events and small for probable ones.

Averaging this quantity over all of the possible states of the system gives a measure of our knowledge about the state of the system

$$H(X) = \sum_{x \in X} p(x) \log \frac{1}{p(x)} = -\sum_{x \in X} p(x) \log p(x)$$

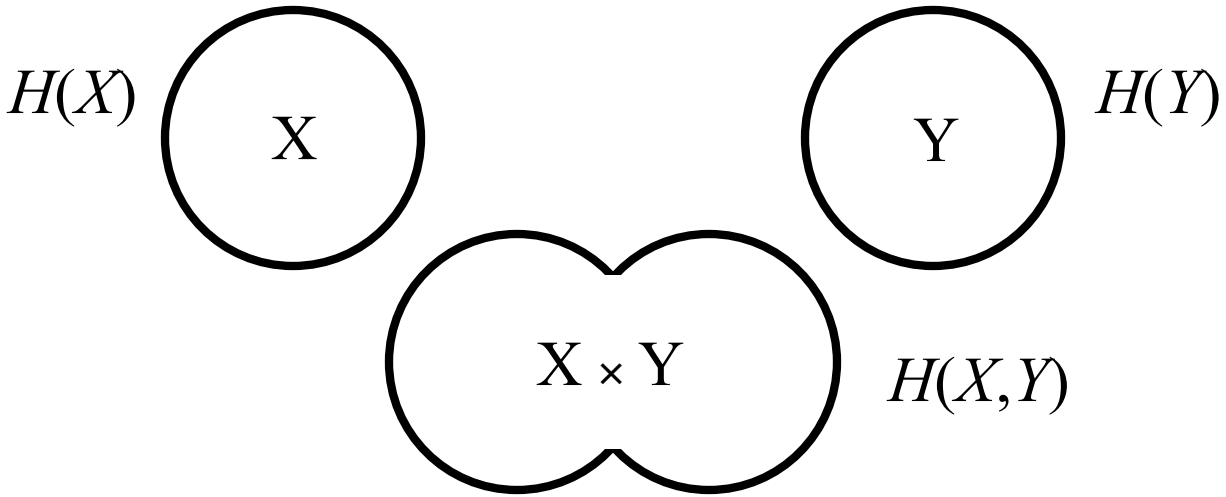
which is called the **entropy**.

Joint Entropy

If the system states can be described with multiple parameters, the entropy is computed by averaging over all possible states

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$

This is called the **Joint Entropy**, since it describes the entropy of the states of X and Y, which jointly describe the system. You can think of X and Y as representing subsystems of the original system $X\times Y$.

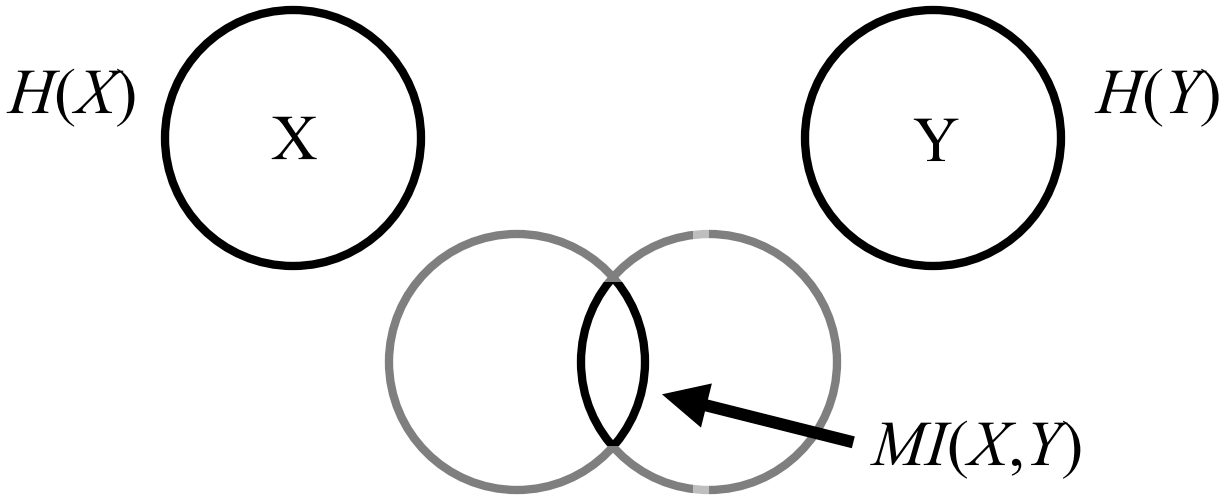


Mutual Information

In this case, an important quantity is the difference of entropies,

$$MI(X,Y) \quad = \quad H(X) + H(Y) - H(X,Y)$$

This is called the **Mutual Information** (MI) since it describes the amount of information that is shared between the two subsystems.



Identifying Relationships

If you know something about subsystem X, the mutual information describes how much information you also possess about Y. For this reason, MI is key in identifying relationships across climate variables, and in identifying and selecting a set of relevant variables that aid in the prediction of another climate variable.

If two climate variables are independent, then the joint entropy is

$$H(X,Y) = H(X) + H(Y)$$

which gives a mutual information of zero, since

$$\begin{aligned} MI(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) + H(Y) - [H(X) + H(Y)] \\ &= 0 \end{aligned}$$

While mutual information can identify dependencies, it cannot determine the causal nature the interaction.

Relation to Probability Densities

The mutual information can also be written as

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$

Which highlights the fact that this is about the **probability density** of the states of the system.

Note that if the joint probability density $p(x,y)$ can be factored into $p(x)\,p(y)$, then the mutual information is zero and the two systems X and Y are independent.

Transfer Entropy

Recently, Schreiber (2000) introduced a novel information-theoretic quantity called the **Transfer Entropy** (TE). Consider two subsystems X and Y, with data in the form of a two time series of measurements

$$X = \{x_1, x_2, \cdots, x_t, x_{t+1}, \cdots, x_n\}$$
$$Y = \{y_1, y_2, \cdots, y_s, y_{s+1}, \cdots, y_n\}$$

then the transfer entropy can be written as

$$T(X_{t+1} \mid X_t, Y_s) \quad = \quad -H(X_t) + H(X_t, Y_s) + H(X_t, X_{t+1}) - H(X_t, X_{t+1}, Y_s)$$

which describes the degree to which information about Y allows one to predict future values of X. This is then a measure of the causal influence that the subsystem Y has on the subsystem X.

Transfer Entropy

We can write this another way using co-informations

$$T(X_{t+1} \mid X_t, Y_s) \quad = \quad I(X_{t+1}, Y_s) - I(X_t, X_{t+1}, Y_s)$$

$$X = \{x_1, x_2, \cdots, x_t, x_{t+1}, \cdots, x_n\}$$
$$Y = \{y_1, y_2, \cdots, y_s, y_{s+1}, \cdots, y_n\}$$

Co-information of Rank 2
(Mutual Information)

$$X = \{x_1, x_2, \cdots, x_t, x_{t+1}, \cdots, x_n\}$$
$$Y = \{y_1, y_2, \cdots, y_s, y_{s+1}, \cdots, y_n\}$$

Co-information of Rank 3

Again, this describes the degree to which information about Y allows one to predict future values of X.

Once can vary s-t, which leads to a TE spectrum.

Estimating Information-Theoretic Quantities

Develop proven tools that will allow researchers to identify relevant variables, and to quantify and characterize their causal interactions.

The basic procedure is straightforward:

Estimating Information-Theoretic Quantities

Develop proven tools that will allow researchers to identify relevant variables, and to quantify and characterize their causal interactions.

The basic procedure is straightforward:

1. Estimate the probability density from which the data were sampled.

Estimating Information-Theoretic Quantities

Develop proven tools that will allow researchers to identify relevant variables, and to quantify and characterize their causal interactions.

The basic procedure is straightforward:

1. Estimate the probability density from which the data were sampled.
2. Using this probability density, estimate the various necessary entropies.

Estimating Information-Theoretic Quantities

Develop proven tools that will allow researchers to identify relevant variables, and to quantify and characterize their causal interactions.

The basic procedure is straightforward:

1. Estimate the probability density from which the data were sampled.
2. Using this probability density, estimate the various necessary entropies.

In practice this is extremely difficult since not only are we interested in the values of these quantities, but we are also interested in the associated uncertainties of our estimates.

We begin by developing optimal histogram models of the probability densities…

Outline

Information Theory

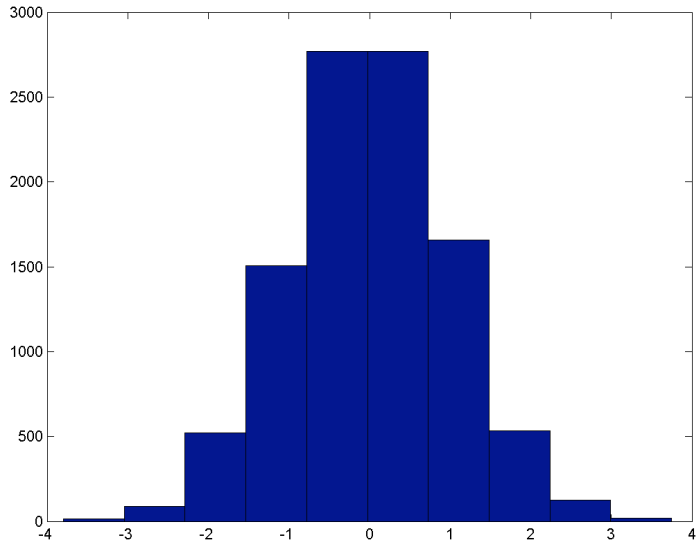
Modeling Probability Densities

Entropy Estimation

Results

Next Steps

Histograms as Probability Density Models

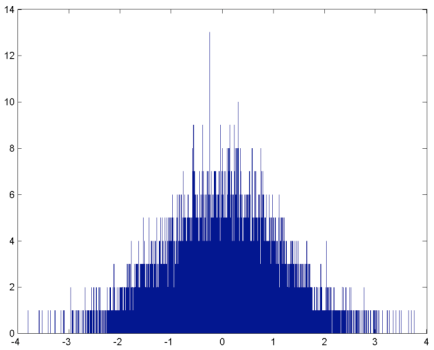


Histograms can be viewed as simple models of the probability density from which the data were sampled.

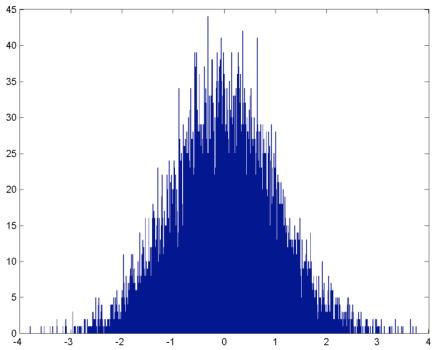
They are convenient since they have regions of constant probability.

Histograms

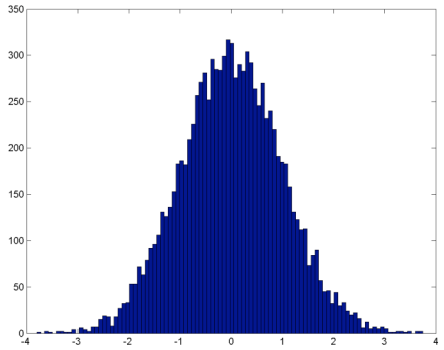
N = 10000, M = 10000



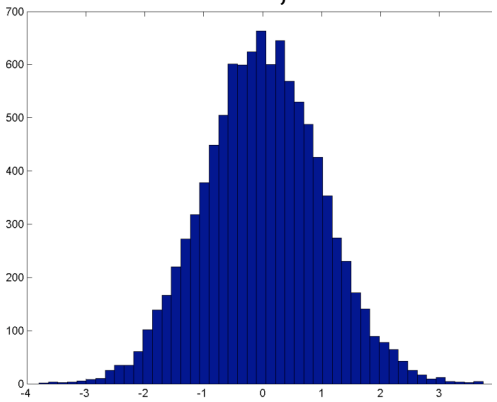
N = 10000, M = 1000



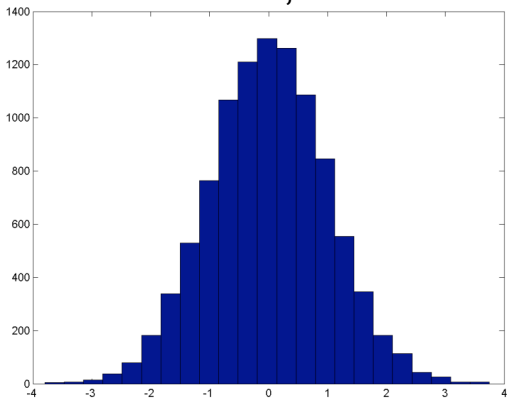
N = 10000, M = 100



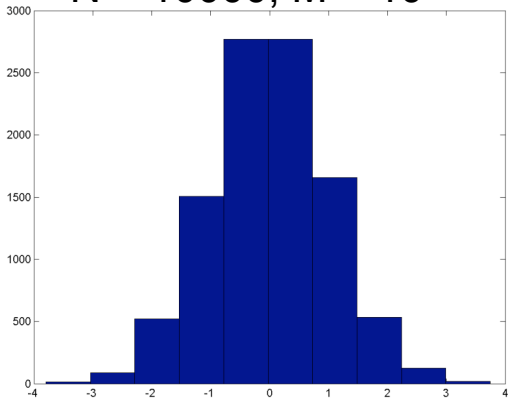
N = 10000, M = 47



N = 10000, M = 23



N = 10000, M = 10



The histogram should contain only details warranted by the data.

Histograms as Probability Densities

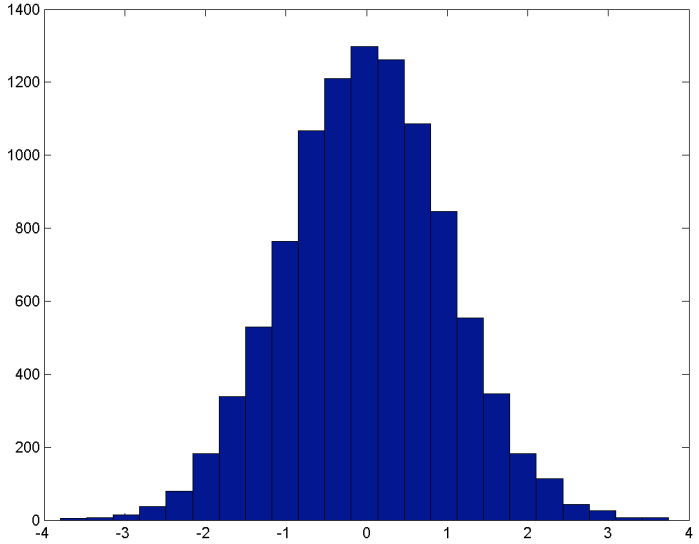
The important concept is that a **histogram is a model** of the underlying probability density from which the data were sampled.

Instead of bins, these are regions of approximate constant probability density.

The result is a constant-piecewise model described by *M* segments, each with probabilities $\pi_1, \pi_2, \pi_3, \ldots, \pi_{M-1}$

We use **Bayesian methods** to find the optimal model parameters

N = 10000, M = 23



Bayes Theorem

Bayes Theorem describes how our prior knowledge about a model, based on our prior information I, is modified by the acquisition of new information or data:

Rev. Thomas Bayes
1702-1761

Likelihood

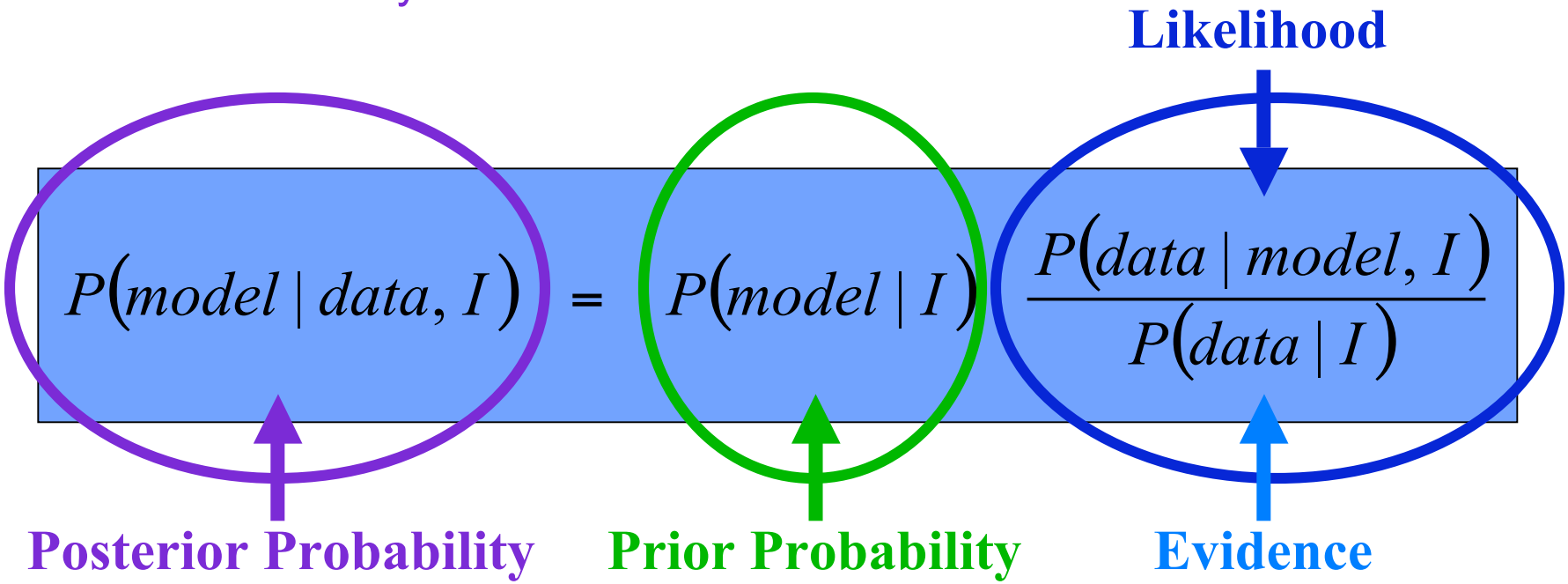
$$P(model \mid data, I) = P(model \mid I) \frac{P(data \mid model, I)}{P(data \mid I)}$$

Posterior Probability **Prior Probability** **Evidence**

Machine Learning

Bayes Theorem is a learning rule.
The Prior Probability describes what you first knew.
Multiply this by a term that describes the effect of your new information, and the result is what you know after you have taken into account your new data.

Rev. Thomas Bayes
1702-1761



Posterior Probability

The joint posterior probability for all the model parameters is

$$p(\partial, M \mid \mathbf{d}, I) \quad \propto \quad \left(\frac{M}{V}\right)^{N} \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{M}} \pi_1^{\,n_1-\frac{1}{2}} \pi_2^{\,n_2-\frac{1}{2}} \ldots \pi_{M-1}^{\,n_{M-1}-\frac{1}{2}} \left(1-\sum_{k=1}^{M-1}\pi_k\right)^{n_M-\frac{1}{2}}$$

It is a product of the priors and likelihood

$$p(\partial, M \mid \mathbf{d}, I) \quad \propto \quad p(M \mid I)\, p(\partial \mid I)\, p(\mathbf{d} \mid \partial, M, I)$$

where

$$p(\partial \mid I) \quad \propto \quad \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{M}} \left[\pi_1 \pi_2 \cdots \pi_{M-1} \left(1-\sum_{k=1}^{M-1}\pi_k\right)\right]^{-1/2}$$

$$p(M \mid I) \quad = \quad constant$$

$$p(\mathbf{d} \mid \partial, M, I) \quad = \quad \left(\frac{M}{V}\right)^{N} \pi_1^{\,n_1} \pi_2^{\,n_2} \ldots \pi_{M-1}^{\,n_{M-1}} \left(1-\sum_{k=1}^{M-1}\pi_k\right)^{n_M}$$

Posterior for the Number of Bins

By integrating over all possible bin probabilities, we can derive the posterior probability of the number of bins given the data.

$$p(M \mid \mathbf{d}, I) \quad \propto \quad \left(\frac{M}{V}\right)^{N} \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{M}} \quad \frac{\prod_{k=1}^{M} \Gamma\left(n_k + \frac{1}{2}\right)}{\Gamma\left(n_1 + b_1 + \frac{3}{2}\right)}$$

It is easier to **find the number of bins that maximizes the logarithm of the posterior probability**

$$\log p(M \mid \mathbf{d}, I) \quad =$$

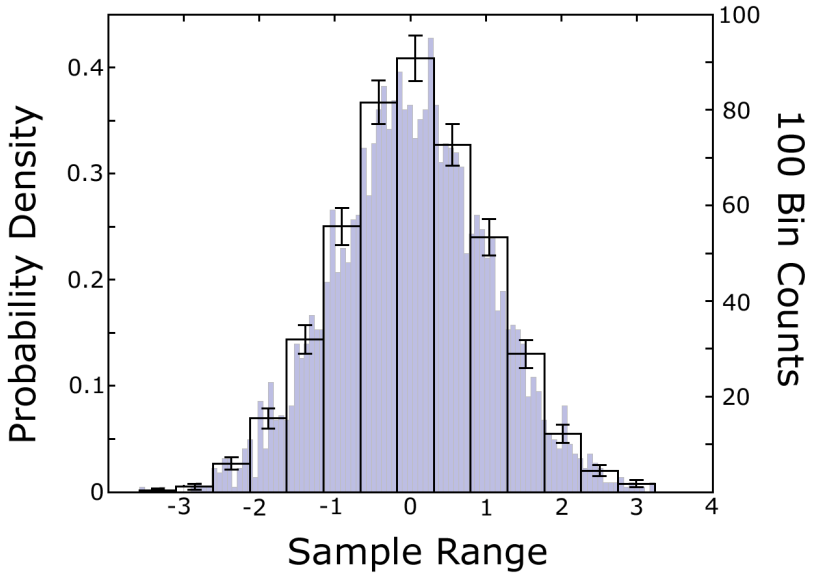
$$N \log M \; + \; \log \Gamma\left(\frac{M}{2}\right) \; - \; M \log \Gamma\left(\frac{1}{2}\right) \; - \; \log \Gamma\left(N + \frac{M}{2}\right) \; + \; \sum_{k=1}^{M} \log \Gamma\left(n_k + \frac{1}{2}\right) \; + K$$

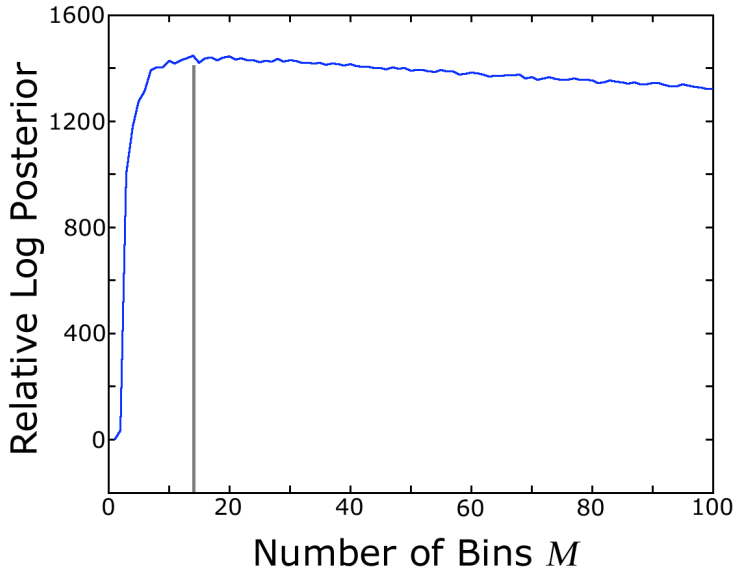
where K is the implicit proportionality constant.

optBINS Algorithm

```
function optM = optBINS(data,minM,maxM)

if size(data)>2 | size(data,1)>1
        error('data dimensions must be (1,N)');
    end

N = size(data,2);


% Loop through the different numbers of bins

% and compute the posterior probability for each.

logp = zeros(1,maxM);

for M = minM:maxM

    n = hist(data,M);  % Bin the data (equal width bins here)

    p = 0;
        for k = 1:M
          p = p + gammaln(n(k)+0.5);
        end

    logp(M) = N*log(M) + gammaln(M/2) - M*gammaln(1/2) - gammaln(N+M/2) + p;

end

[maximum, optM] = max(logp);

return
```

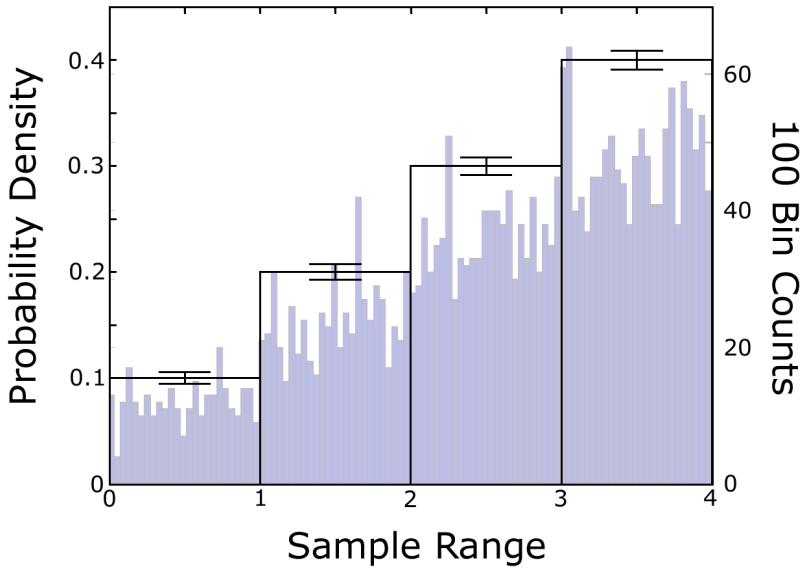
Optimal Histograms

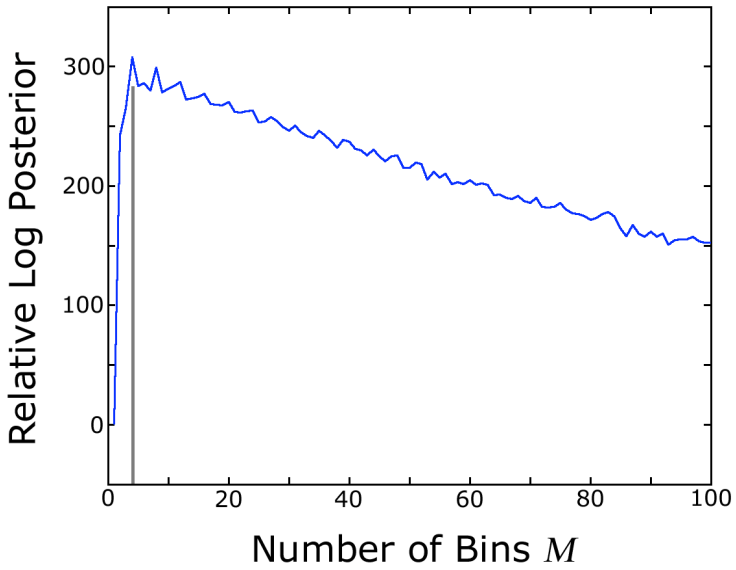




Optimal Binning for N = 3000 Gaussian distributed data points: **M = 14**

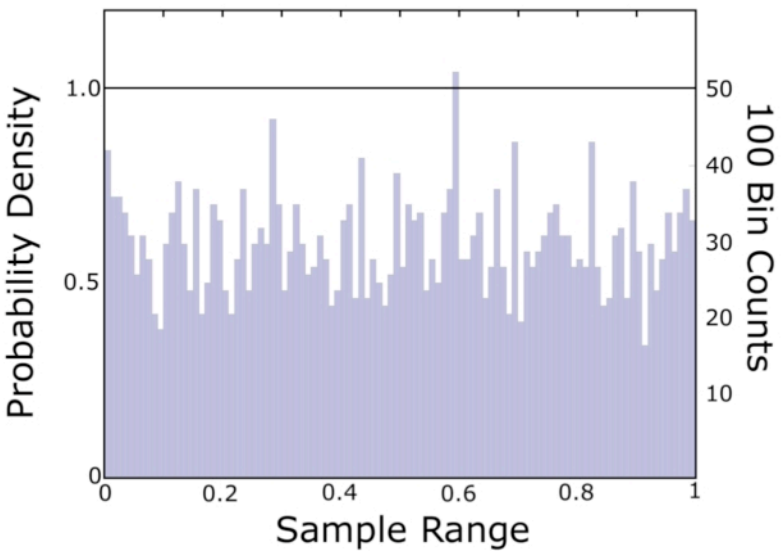
Optimal Histograms

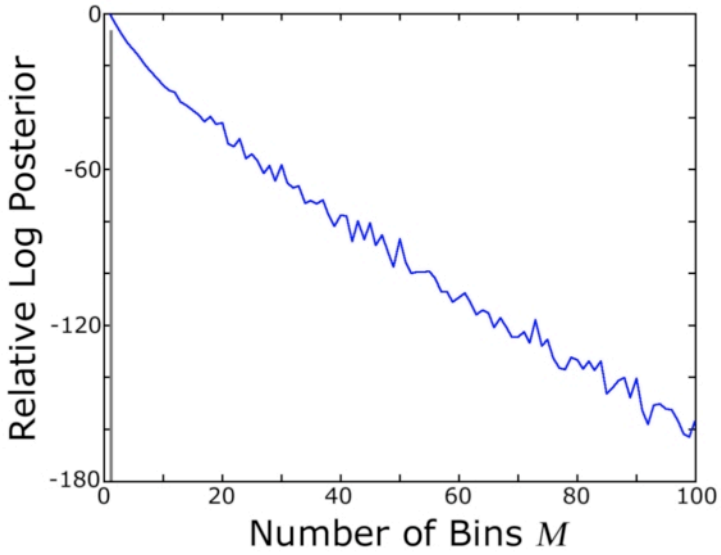




Optimal Binning for N = 3000 data points from a staircase density: **M = 4**

Optimal Histograms

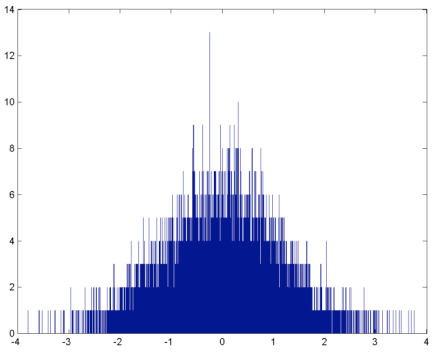




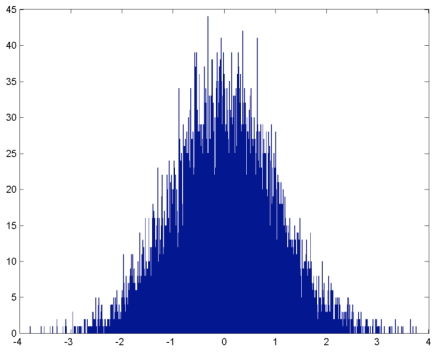
Optimal Binning for N = 3000 data points from a uniform density: **M = 1**

The Optimal Histogram

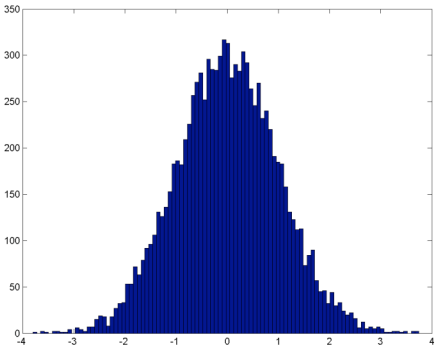
N = 10000, M = 10000



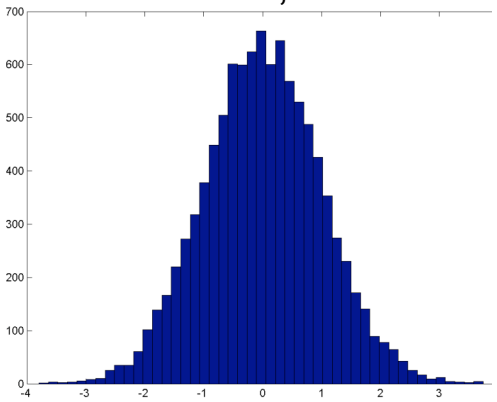
N = 10000, M = 1000



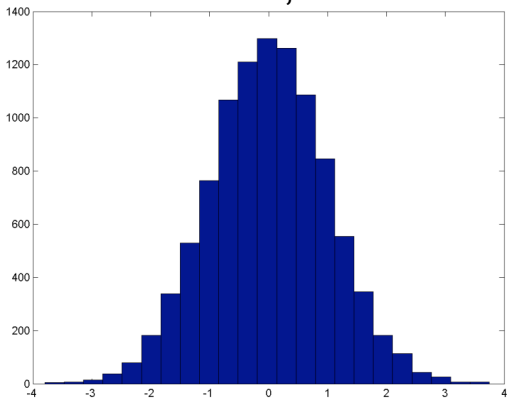
N = 10000, M = 100



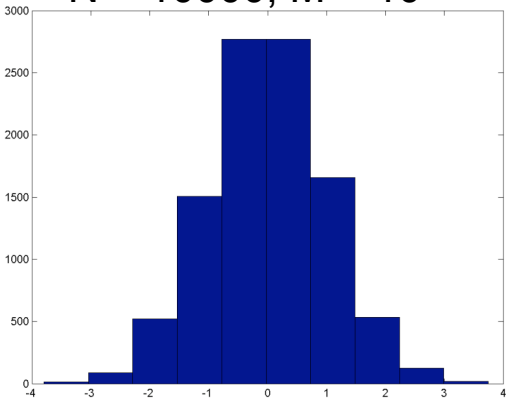
N = 10000, M = 47



N = 10000, M = 23

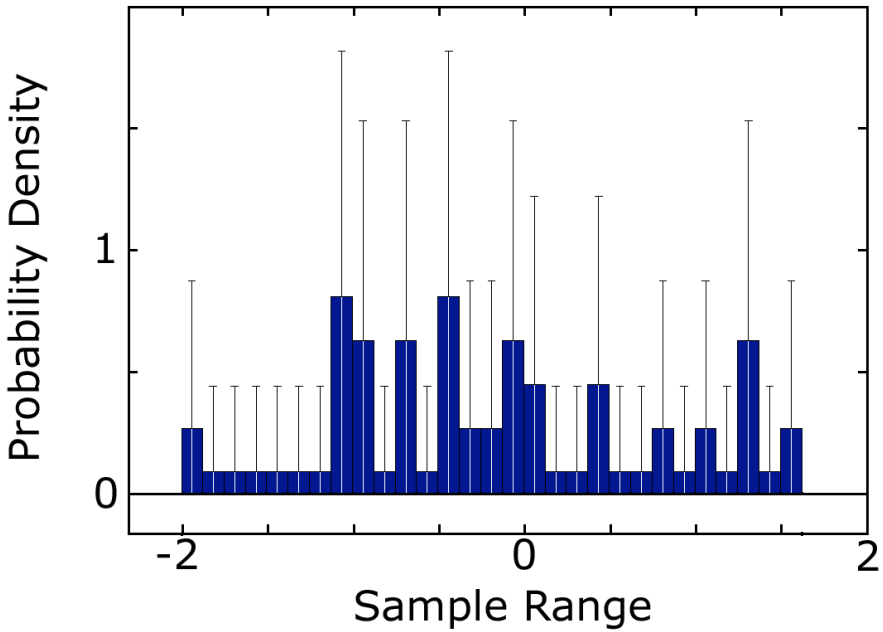


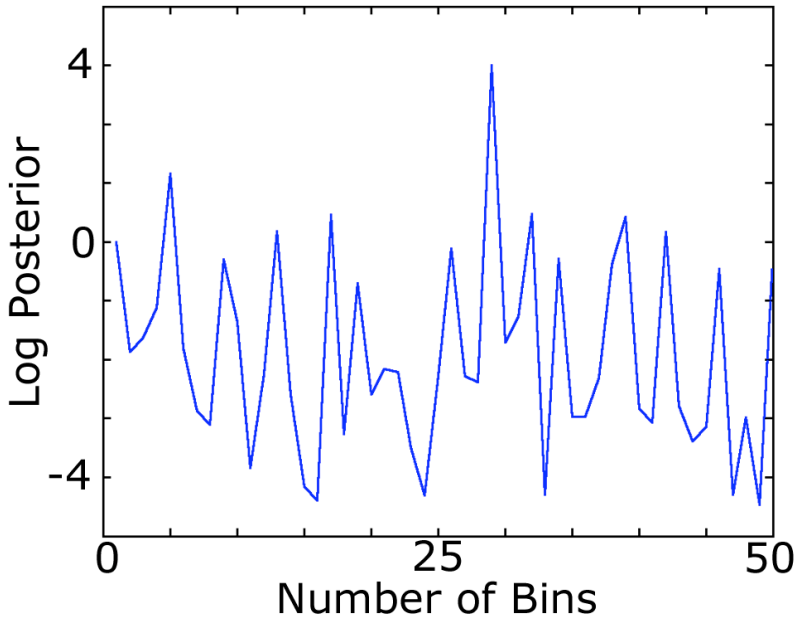
N = 10000, M = 10



The histogram should contain only details warranted by the data.

Sufficient Data?



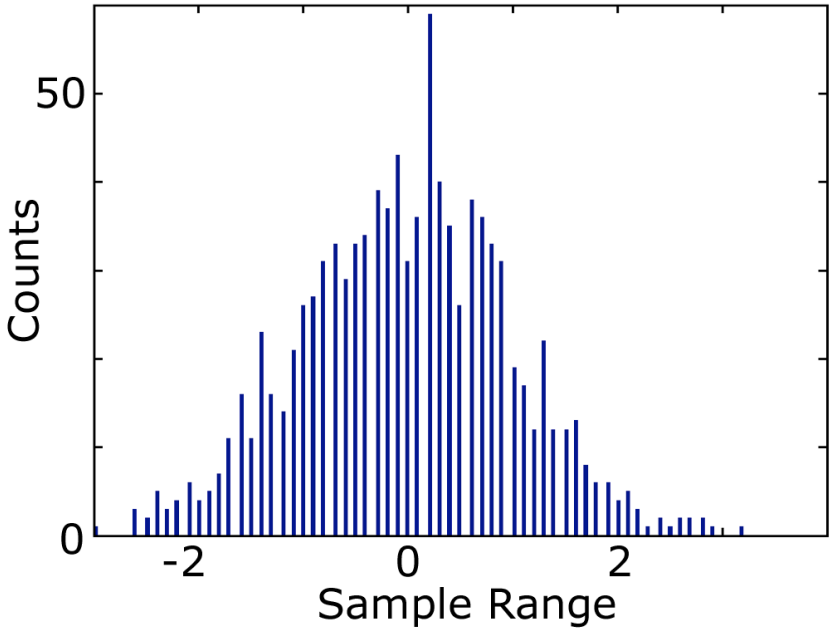


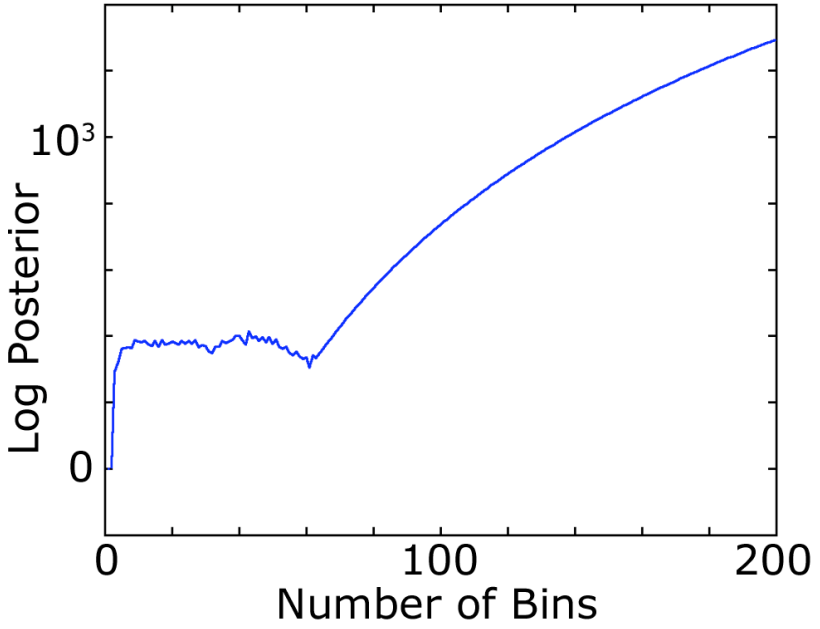
When there are insufficient data (N = 30), the posterior probability has multiple local maxima, and the error bars on the bin probabilities are large.

None of the bin heights are known to be significantly greater than zero.

We have found that 100-150 data points are necessary to estimate a probability density.

Excessive Round-Off?





In the case where the data have been excessively rounded, the discrete nature of the data is a relevant feature.

In this case, information has been lost and cannot be recovered.

Error Bars on Density Parameters

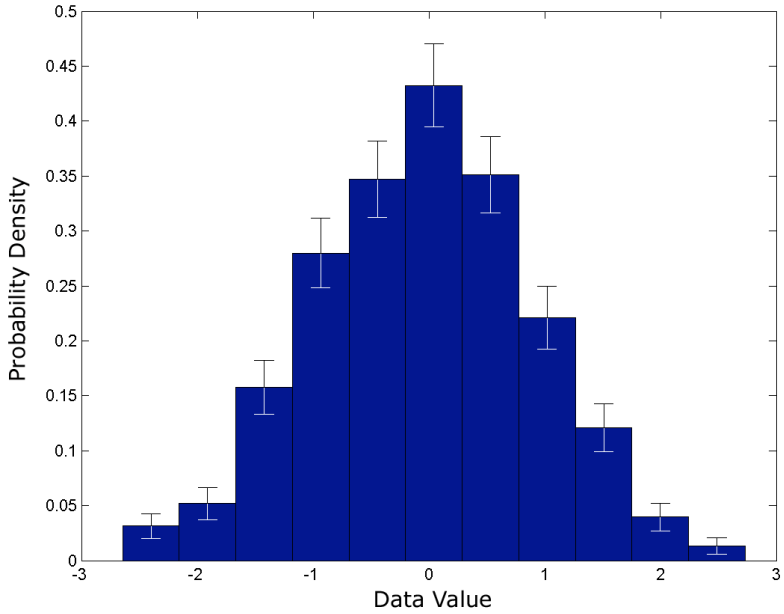
The posterior probability can be used to compute the mean bin probabilities as well as their standard deviations.

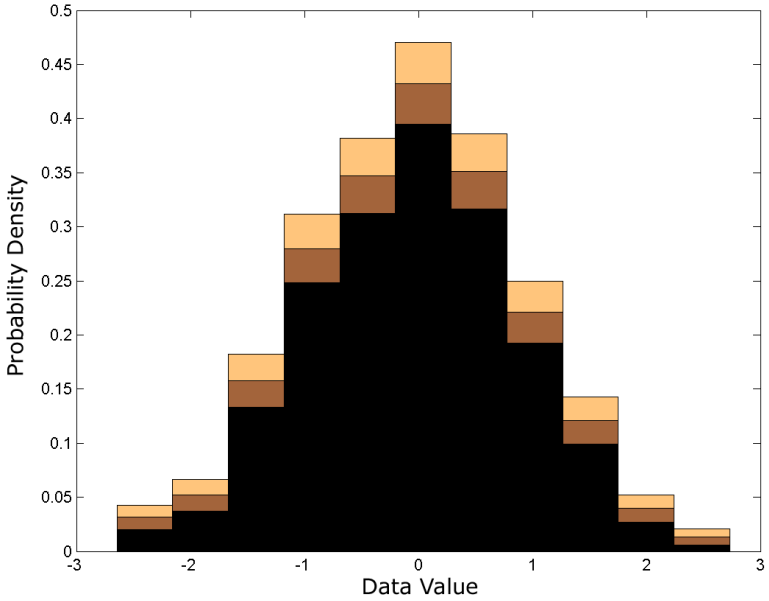
$$\langle \pi_i \rangle = \frac{n_i + \frac{1}{2}}{N + \frac{M}{2}} \qquad \sigma_i^2 = \frac{\left(n_i + \frac{1}{2}\right)\left(N - n_i + \frac{M-1}{2}\right)}{\left(N + \frac{M}{2} + 1\right)\left(N + \frac{M}{2}\right)^2}$$

Where N is the number of data points, M is the number of bins, n_i is the number of data points in the i^{th} bin.

Note that a bin still has a nonzero probability even if there are no counts. Just because you don't have data for an event doesn't rule out that the event cannot occur. Bayes handles this naturally.

Probability Density Visualization



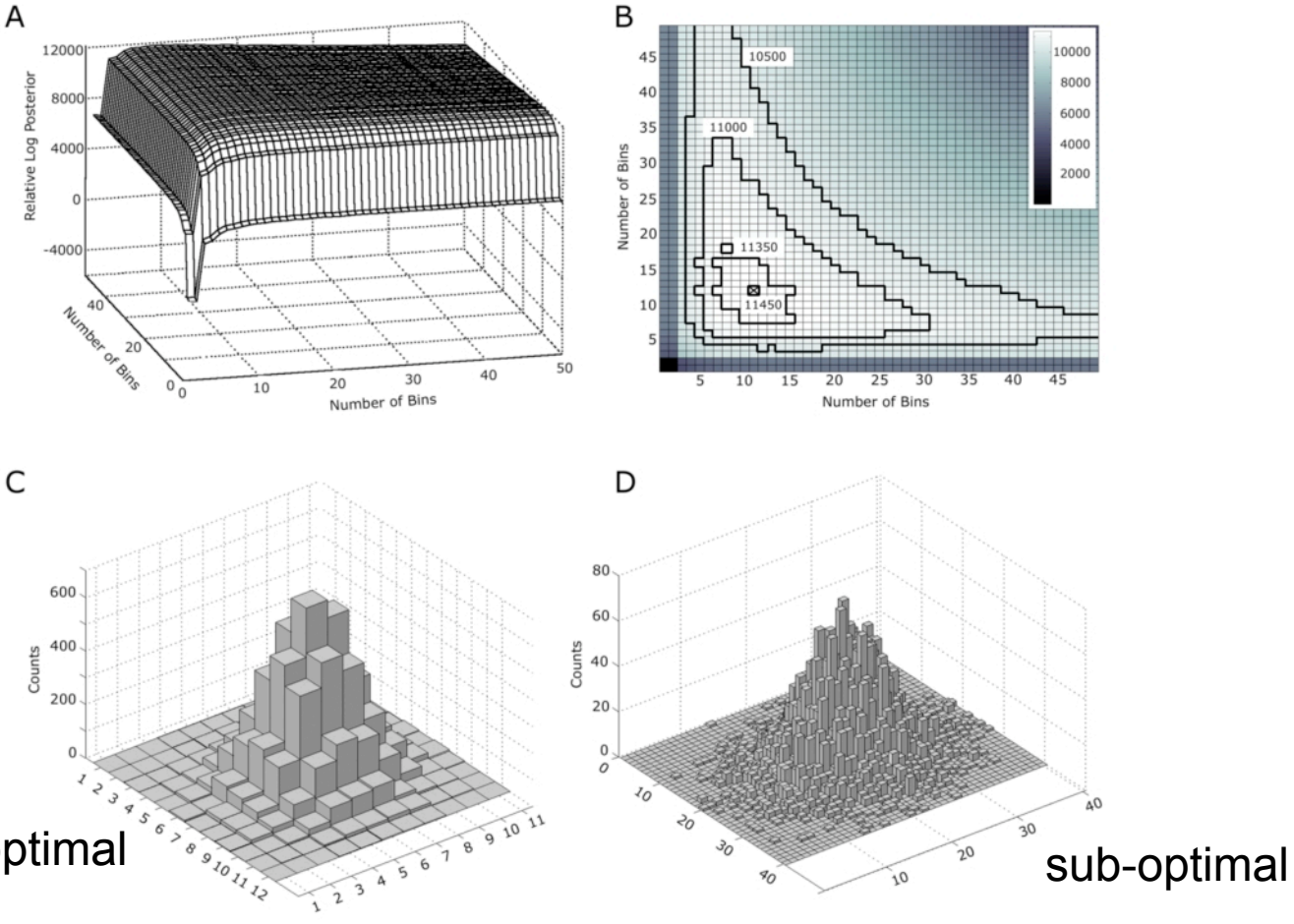


We have written two `Matlab` visualization routines that display histogram models of probability density functions along with the error bars. These error bars signify the inherent uncertainties in our inferences from a finite data set. The heights of the bars represent the mean density and the error bars represent the standard deviations about the mean.

Extendable to Multi-Dimensional Densities

N = 10000

M = 11,12



Extendible to multi-dimensional histograms

Outline

Information Theory

Modeling Probability Densities

Entropy Estimation

Results

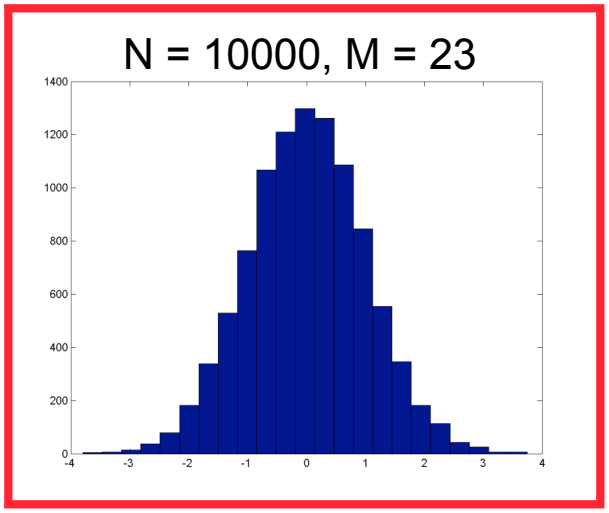
Next Steps

Entropy Estimation

Entropy estimation is relatively easy with a constant-piecewise model

$$H = -\sum_i p_i \log p_i$$

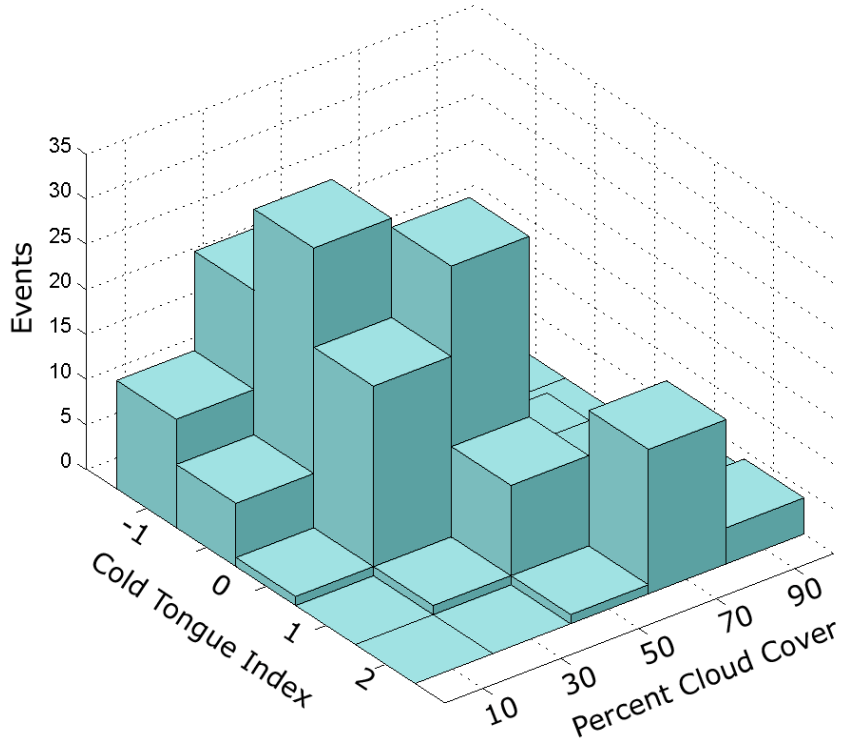
```
H = -sum(p .* (log(p) - log(vol)));
```



Entropy Estimation

And also in higher-dimensions…

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$



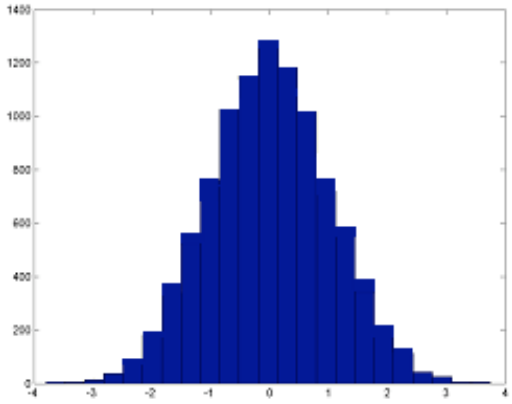
How to Obtain the Uncertainties

To calculate the uncertainties in the entropy estimates, one must first realize that we are uncertain as to the bin probabilities of the probability density model.

By sampling a set of bin probabilities, we obtain a set of probable density functions, along with a set of probable entropies.

$$p(\partial, M \mid \mathbf{d}, I) \quad \propto \quad \left(\frac{M}{V}\right)^{N} \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{M}} \pi_1^{\,n_1-\frac{1}{2}} \pi_2^{\,n_2-\frac{1}{2}} \ldots \pi_{M-1}^{\,n_{M-1}-\frac{1}{2}} \left(1-\sum_{k=1}^{M-1}\pi_k\right)^{n_M-\frac{1}{2}}$$

From this set of probable entropies, we can compute the mean and variance. Thus quantifying both the entropy and our uncertainty.



Entropies from Sampling

This shows some of the results from sampling from the posterior probability and computing the entropies.

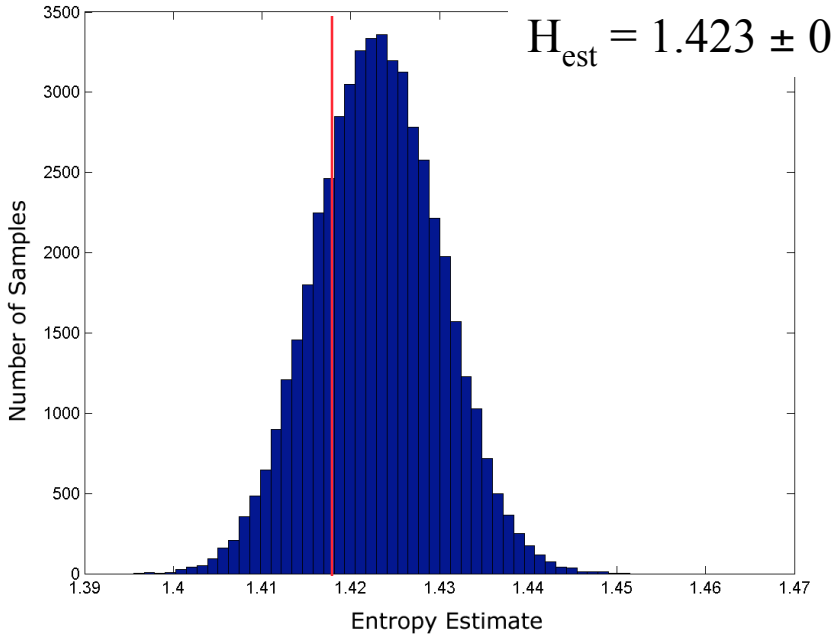
The data was from a Gaussian distribution with $\mu = 0$, $\sigma = 1$.

The true entropy is $H_{true} = 1.419$

$N = 10000$, $M = 24$

50000 Samples

H = 1.4202
1.4161
1.4159
…
1.4211
1.4259
1.4290



Outline

Information Theory

Modeling Probability Densities

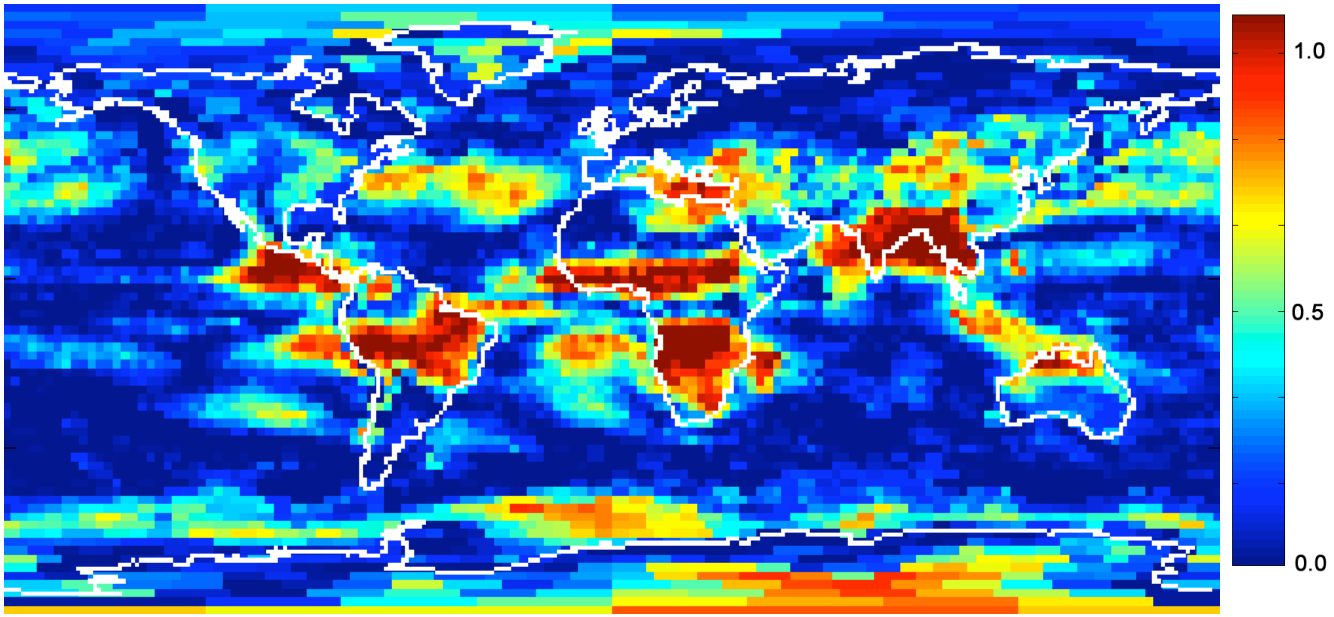
Entropy Estimation

Results

Next Steps

Preliminary Mutual Information Results

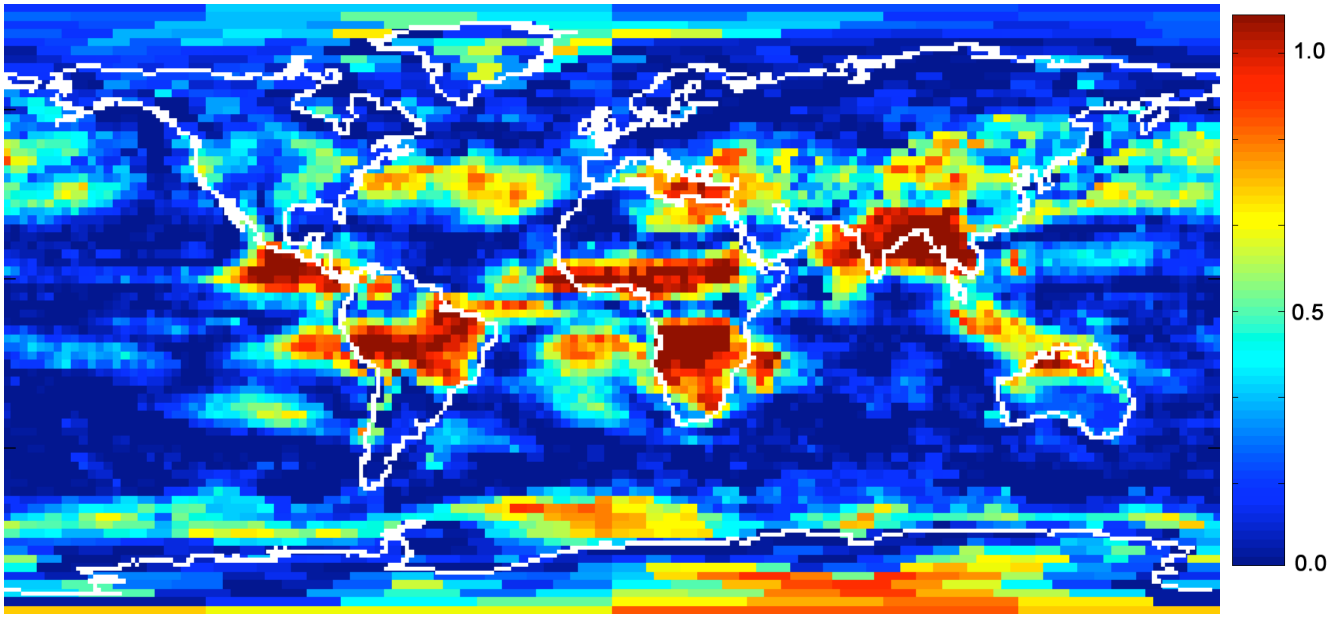
Mutual Information between ISCCP percent cloud cover and Seasonality.



The data consisted of monthly averages of percent cloud cover resulting in a time-series of 198 months of 6596 equal-area pixels each with side length of 280 km. The analysis was performed pixel-wise so that for each pixel:
X = cloud cover percentages and Y = month of the year (seasonal state).
The MI was computed for each pixel independently and is color-coded on the map above.

Cloud Cover and Seasonality

Mutual Information between ISCCP percent cloud cover and Seasonality.

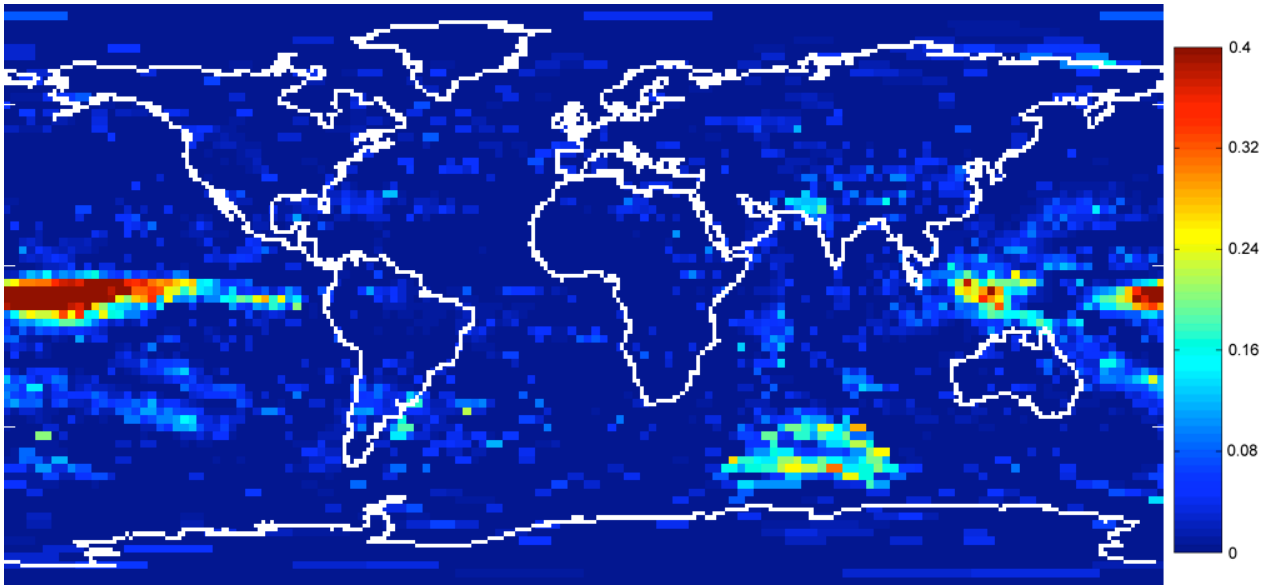


This method finds the Inter-Tropical Convection Zones, The Monsoon Regions, the Sea Ice off Antarctica, and cloud cover in the North Atlantic and Pacific.

This figure can be directly compared to the PCA analysis performed by Rossow et al. 1991, J. Climate, 6:2394-2418.

Cloud Cover and Cold Tongue Index

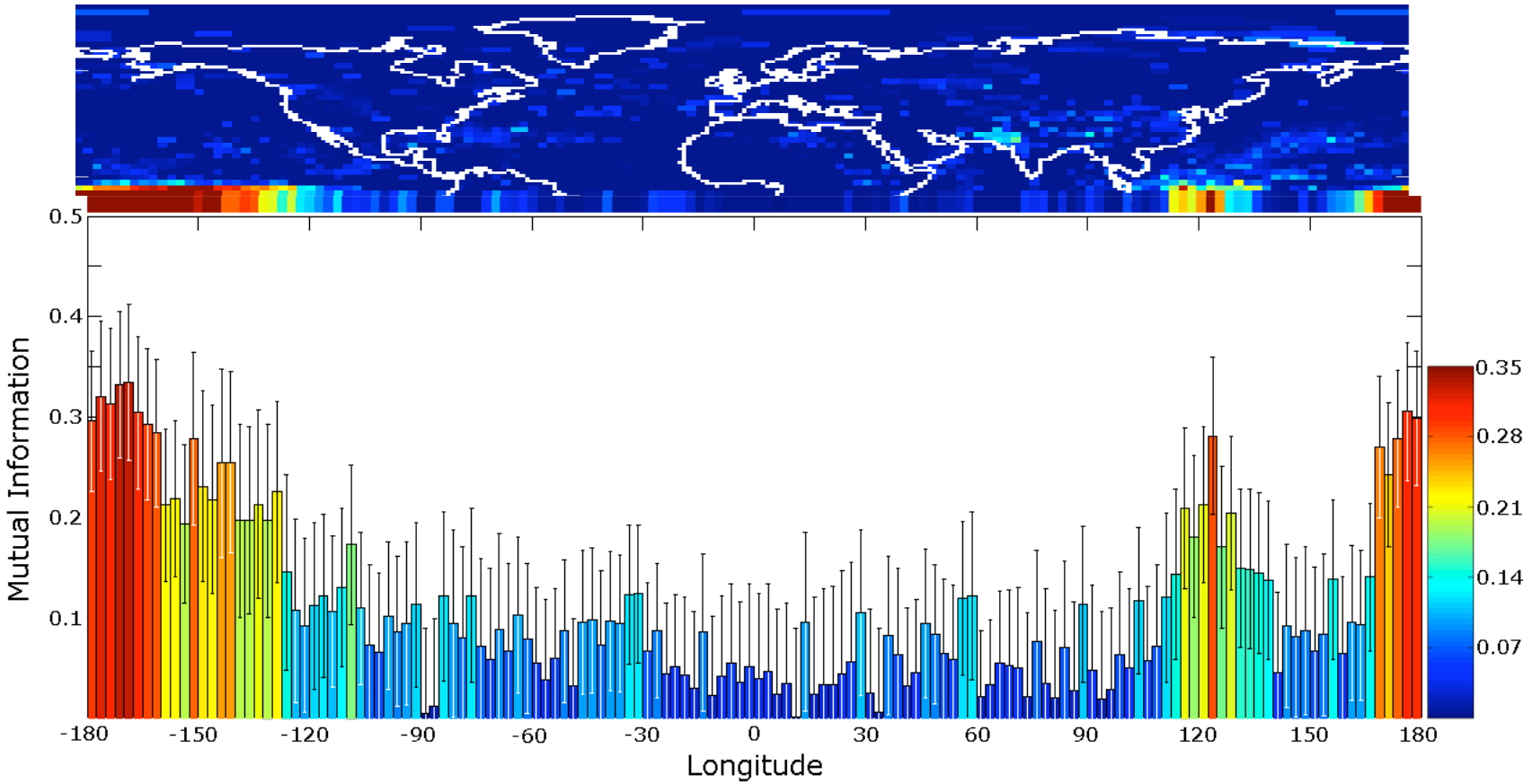
Mutual Information between ISCCP percent cloud cover and Cold Tongue Index (CTI), which describes the sea surface temperature anomalies in the eastern equatorial Pacific Ocean (6N-6S, 180-90W) and is indicative of ENSO variability.



Cloud cover data is from ISCCP C2 and CTI data is from T. Mitchell:
http://tao.atmos.washington.edu/pacs/additional_analyses/sstanom6n6s18090w.html

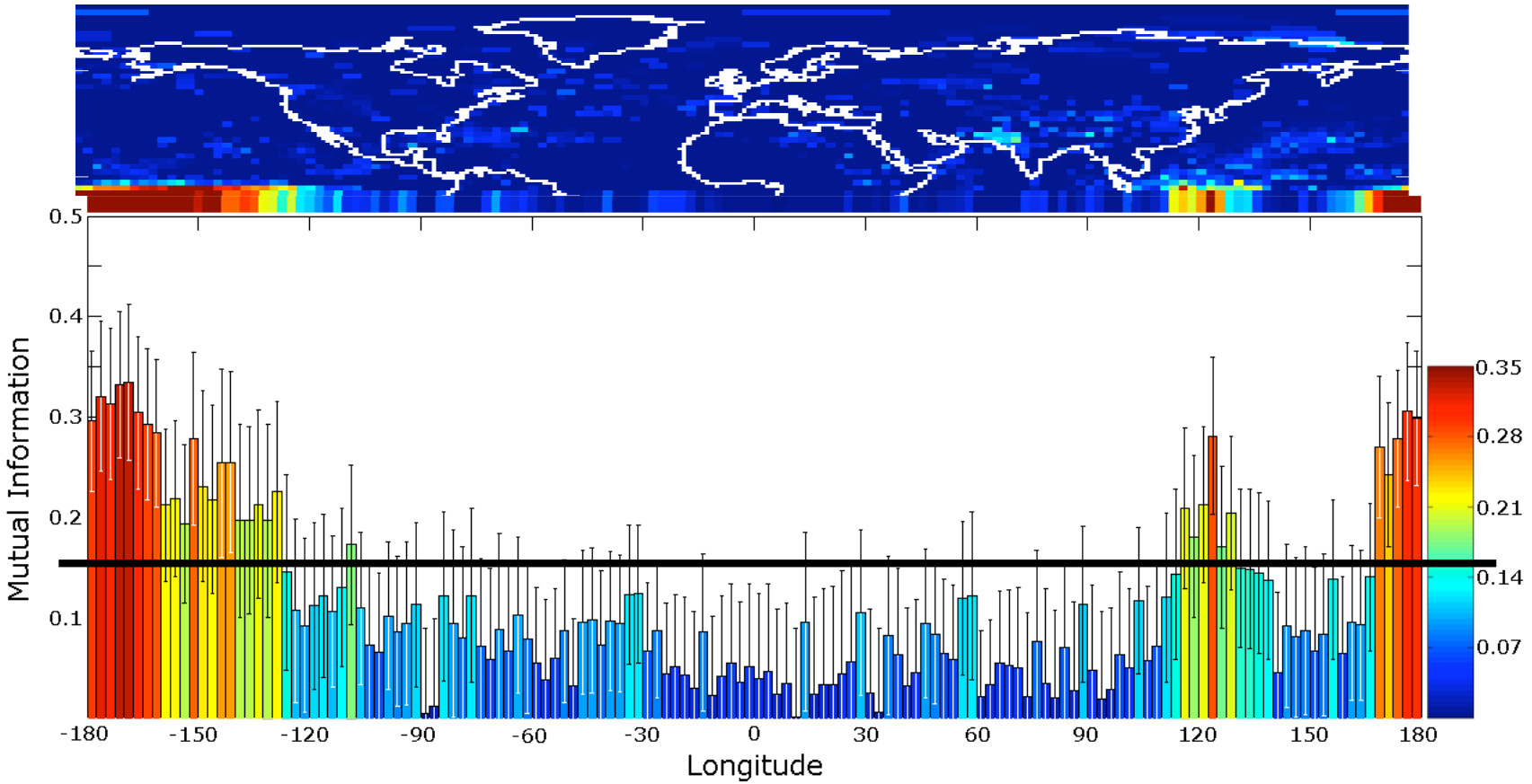
New Refined Results

Refined analysis with error bars



Statistical Significance

Values below line do not indicate statistically significant interactions



Outline

Information Theory

Modeling Probability Densities

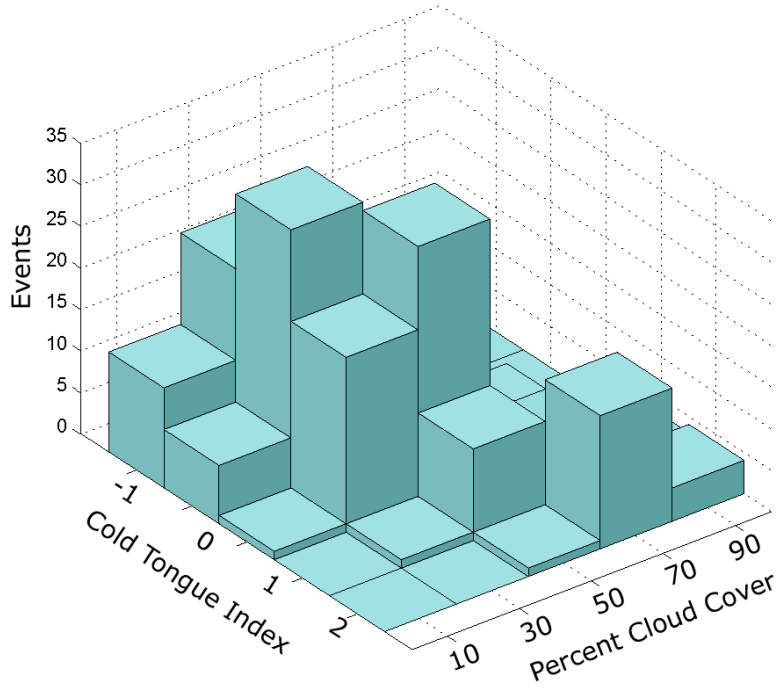
Entropy Estimation

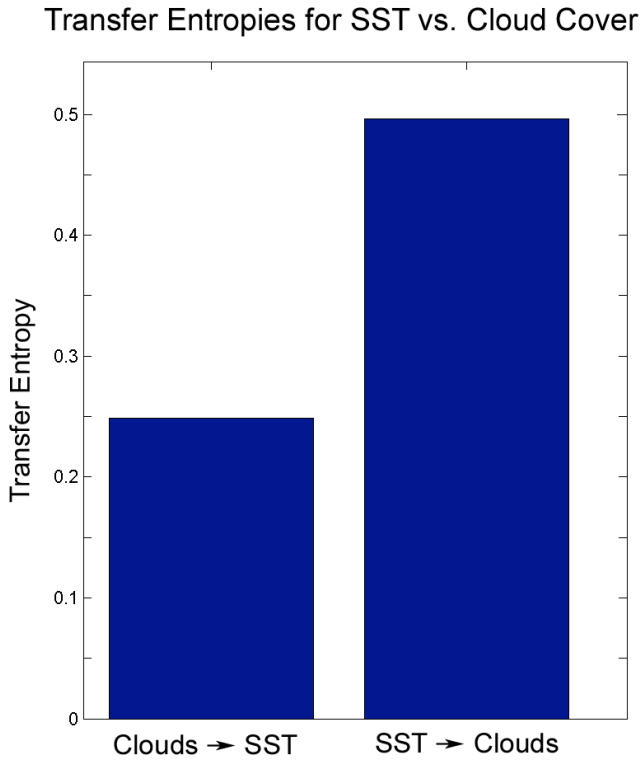
Results

Next Steps

Transfer Entropy

We are currently completing code to compute Transfer Entropies with Error Bars. Our earlier results indicate that this is useful as a potential indicator of causal interactions.





Data from 1.25° N 191.25° W

Thanks to the NASA ESTO AIST Program for their support